POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Analysis of Milan Personal Services - Database 12

Group Project Report in DATA INFORMATION AND QUALITY (2024-2025)

Mauro Orazio Drago, 10947649
Dennis Pierantozzi, 10959535
Davide Morelli, 10962902

---

1. Introduction

The aim of this report is to present our project for the "Data and Information Quality" exam. The project requested to execute a complete Data Preparation Pipeline on the dataset that has been assigned to our group following precise steps. Specifically, we conducted:

- Data Profiling and Assessment
- Data cleaning
- Data Analysis.

1.1. Set up

The dataset pertains to beauty services in Milan and is openly available at DatiOpen.it. It contains various details about each shop, including address, type of exercise (e.g., hairdresser, barber, beautician), main activity, and area in square meters.
We used Pandas, NumPy, Sklearn, Rapidfuzz and Scipy libraries for processing and Kaggle as our development environment. To maintain a clear structure, we divided our code into three notebooks, which are reflected in the section of this report.

2. Data Assessment

The dataset consists of 3909 rows and 10 columns:

- "Tipo esercizio pa": type of work performed (e.g., hairdresser, beautician)
- "Ubicazione": full address (type of public space, name of the public space, address number, municipality number)
- "Tipo via": type of public space
- "Via": name of the public space
- "Civico": address number
- "Codice via": unique identifier of the public space
- "Zd": municipality number of Milan
- "Prevalente": main activity carried out
- "Superficie altri usi": square meteres not used directly for the business
- "Supericie lavorativa": square meteres used directly for the business

2.1.

Data Types and Initial Observations:

- Most columns are of type "object," except for "Codice via," "Superficie altri usi," and "Superficie lavorativa," which are float64.
- The dataset includes one duplicate row.

2.2. Data Quality Dimensions:

Data Assessment aims to define a score for the data quality dimensions chosen. This can be done in a subjective or objective way. The subjective method requires the usage of questionnaires to assess the scores through the answers gathered. This process requests also a final interpretation step.
The objective way, instead, measures the scores using mathematical tools and formulas. Because of temporal limits, we have decide to perform an objective assessment only.

We assessed completeness, consistency, and accuracy but excluded timeliness due to the absence of temporal metadata.

2.2.1 Completeness

Completeness is the degree to which a given data collection includes the data describing the corresponding set of real-world objects. Ojbectively the completeness is computed as the total number of not null values divided by the total number of cells.

The completeness of our dataset is 79%.
The columns that contains the major number of null values are "Prevalente", "Superficie altri usi".

2.2.2 Consistency

Consistency is the satisfaction of semantic rules defined over a set of data items. Considering our dataset we have defined some specific rules that we have checked.

- Rule 1: The "Ubicazione" column must equal the concatenation of "Tipo via," "Via," "Civico," and "Zd."
  Issues found:
  - 152 mismatches with "Civico"
  - 21 mismatches with "Tipo via"
  - 88 mismatches with "Via"
  - 18 mismatches with "Zd"
- Rule 2: Values in "Superficie lavorativa" and "Superficie altri usi" must be non-negative.
  Result: no negative values were found.

2.2.3 Accuracy

Accuracy dimension check the extent to which data are correct, reliable and certified. There are two types of accuracy: the syntactic and semantic accuracy. The former checks whether or not a value is correct syntactically with respect of a domain of allowed values. Semantic accuracy checks whether or not a value has a real meaning in the real world.

- **Semantic Accuracy**: We have verified the validity of "Tipo via" and "Via" against external datasets (TIPOVIA and VIARIO).
  - "Tipo via": 99.9% accuracy (3905 correct out of 3908 non-null values).
  - "Via": 97.2% accuracy (3800 correct out of 3908 non-null values).
- **Syntathic Accuracy**: We have checked if the values in column "Zd" are in the number of municipality of Milan (1-9).
  - "Zd": 99.97% accuracy (2 invalid values out of 3909).

2.3. Data Profiling

Data Profiling is the set of activities and processes designed to determine the metadata of a given dataset.

We have performed a single column analysis per each column that has discrete values, inspecting the number of unique values for each of them and counting the most and least common ones while for the continuous columns "Superficie altri usi" and "Superficie lavorativa" we have studied the value distributions.

2.3.1 Single Column Analysis

- "Tipo esercizio pa": 103 unique values (most frequent: "Parrucchiere per signora," least frequent: "Truccatore").
- "Ubicazione": 3554 unique values.
- "Tipo via": 17 unique values (most frequent: "VIA," least frequent: "SIT").
- "Via": 1370 unique values.
- "Civico": 235 unique values.
- "Codice via": 1376 unique values.
- "Zd": 10 unique values (1-9).
- "Prevalente": 63 unique values.

2.3.2 Value distribution

- "Superficie altri usi": Mean = 9.56, Std Dev = 13.73
- "Superficie lavorativa": Mean = 40.15, Std Dev = 27.73

Below are reported two visualization (e.g., histograms and boxplots). They demonstrate positive skewness for both fields, suggesting that median imputation is preferable for missing values.

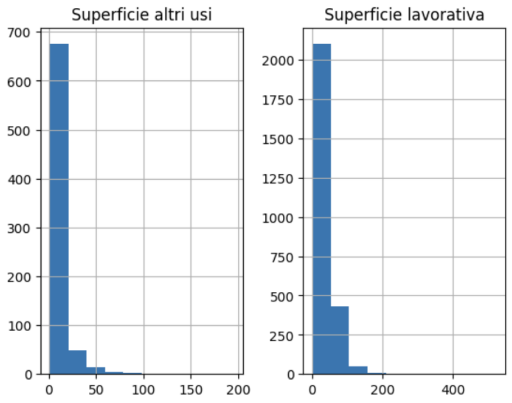

Figure 1: Caption

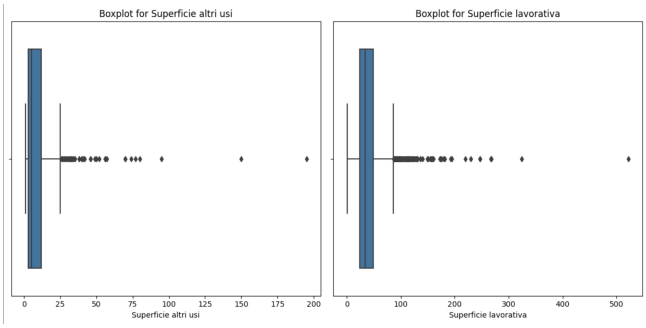

Figure 2: Caption

3. Data Cleaning

Data cleaning si the process of identifying and eliminating inconsistencies, discrepancies and errors in data in order to improve quality.

3.1. Data Wrangling

3.1.1 Column Renaming

To improve readability and data management, the dataset columns were renamed using a standardized format: lowercase names separated by underscores. This facilitates access and manipulation of the data during subsequent analysis phases. Specifically:

- Tipo esercizio pa → t_es (business type)
- Ubicazione → ubic (location)
- Tipo via → t_via (street type)
- Via → via (street name)
- Civico → civ (street number)
- Codice via → codice_via (street code)
- ZD → zd (zone designation)
- Prevalente → main (primary activity)
- Superficie altri usi → sup_alt (surface area for other uses)
- Superficie lavorativa → sup_lav (working surface area)

3.2. Data Transformation

Data transformation involved several operations to standardize and enrich the dataset, aiming to prepare it for more in-depth analysis.

3.2.1 Correction of "zd" values

Inconsistent values in the "zd" column were identified and corrected. In particular, rows representing the same object were merged, ensuring the consistency of "zd" values

3.2.2 Reduction of Labels in "t_es"

The "t_es" column contains 103 unique labels. Many of these labels have similar meanings or represent variations of the same concept, leading to redundancy and potential inconsistencies in analysis. To address this, we will reduce the number of labels by grouping similar ones under a unified label. This will make the dataset easier to interpret and analyze while preserving the essential distinctions between categories.
To achieve this, we first exploded the labels separated by the character ";" and applied composite label handling and label mapping based on a set of labels that we identified.
In particular, we studied the laws governing personal services in Italy. Beauty services are categorized into different types, as represented in the dataset, and we have maintained these distinctions. The "Acconciatore" and "Parrucchiere" labels were merged into a single category following a law passed in 2018. Since our dataset collects data from 2012, we have decided to retain this distinction in the dataset.
At the end we have:

- 907 Acconciatore
- 601 Parrucchiere per Uomo
- 1294 Parrucchiere per Donna
- 1266 Tipo A - Estetica Manuale
- 693 Tipo B - Centro di Abbronzatura
- 226 Tipo C - Trattamenti Estetici Dimagrimento
- 162 Tipo D - Estetica Apparati Elettromeccanici
- 29 Centro Benessere
- 139 Centro Massaggi
- 24 Esecuzione di Tatuaggi e Piercing
- 1 Truccatore

3.2.3 Label Reduction for "main" column

The "main" column contains many null values, and the few non-null values it holds are not significantly different from the information already present in the t_es column. Therefore, we use the values from the main column to populate the t_es column where t_es is null, ensuring that all relevant information is maintained in a single column.

3.2.4 Format Verification

- "t_via": contains an invalid type "Vie". We have mapped this value to the type "Via", which is one of the types allowed in our dataset, TIPOVIA, that we have taken as ground truth.
- "cod_via": all values were integers, so no changes were made.
- "civ": We ensured that all values followed the expected format: 1-3 digits, optionally followed by an uppercase letter (e.g., 12A), a slash "/" with a single digit (e.g., 93/1), or a slash "/" and another series of digits and letters (e.g., 15/101). Invalid values have been replaced with "NaN", ensuring that only correctly formatted entries remain for further processing.
- When processing records, some entries in the via column contained redundant information that matched part of the t_via column. For example, a value for "via" was CSO COMO, which included the redundant prefix CSO. To ensure consistency and follow a specific order, we removed the redundant portion.
  We then validated the values against those in the VIARIO dataset. Similarity scores were calculated between the via column and two descriptive fields, using the highest score to flag rows for review if they fell below a 75% similarity threshold. The similarity measure used in this case was the edit distance, implemented using the RapidFuzz library.

3.2.5 "Ubic" Structuring and Format Checking

We processed the ubic field to extract and organize its contents into distinct components: ubic_t_via (street type), ubic_via (street name), ubic_civ (street number), ubic_zd (zone description), and note (additional information).
Next, we standardized the values and validated them against the VIARIO dataset by calculating similarity scores between the ubic_via column and the DESCRITTIVO and DENOMINAZIONE columns of the external dataset. The highest similarity score was selected for each row, and rows with a similarity score below 75% were flagged for further review. For the rows that did not have a similarity score higher than 75%, we performed a second-level review using the Jaccard distance. With this similarity measure, most of the rows now exhibit a similarity of 50% or greater, indicating that their addresses match semantically. As a result, the ubic_via column can now be standardized to align with the format of the DESCRITTIVO column of the VIARIO dataset.

3.3. Error Detection and Correction

In this section we aim to solve the mismatches between the the ‘ubic‘ columns (‘ubic_t_via‘, ‘ubic_via‘, ‘ubic_civ‘, ‘ubic_zd‘) and the corresponding main columns (‘t_via‘, ‘via‘, ‘civ‘, ‘zd‘). Mismatches are identified by comparing values in these columns.
The mismatches were:

- 20 for "t_via"
- 151 for "civ"
- 76 for "via"
- 17 for "zd"

Per each column we have studied the mismatches and clean the data. For the Zd mismatches we have used the dataset VIARIO to assesses the true values.

4. Data Analysis

In this section, we performed data analysis on both the raw and cleaned datasets. We focused on a classification task for the "Tipo esercizio pa" column, aiming to predict its value based on the other available data.

4.1. Dirty Dataset Pipeline

We selected only the columns likely to impact performance, discarding the others.
The features we retained are:

- "Tipo esercizio pa"
- "Tipo via"
- "Via"
- "Zd"
- "Superficie altri usi"
- "Superficie lavorativa"

4.1.1 Encoding

Some features were categorical, so we applied encoding methods. Since "Tipo esercizio pa" had 103 unique values, we used Label Encoding with the `LabelEncoder` from the `sklearn` library. For the categorical columns "Tipo via" and "Zd," we applied the one-hot encoding technique.

4.1.2 Handling Null Values

Rows with null values in "Tipo esercizio pa" were dropped. For "Superficie lavorativa" and "Superficie altri usi," missing values were filled with the median of the existing values.